

Shu Yang, Ph.D.

Assistant Professor of Statistics
North Carolina State University
E-mail: syang24@ncsu.edu

NCSU-Statistics Department
2311 Stinson Drive
Campus Box 8203
Raleigh, NC 27695-8203

Research Interests

Many important questions in chronic diseases and cancer are about the effects of treatments, e.g. approving drugs, implementing health policies, or identifying optimal personalized treatment strategies. The answers to these questions often rely on complex real-world data suffering from confounding, non-compliance, drop-outs, missing values, and etc.

My research is to develop innovative statistical methods for making accurate inferences about treatment effects from complex observational and clinical studies, including marginal structural models, structural nested models, inverse probability weighting, and matching methods. This research falls into the general area of causal inference and missing data analyses. Our research team applies the novel methods in cardiovascular diseases, HIV infection and cancer research to identify effective treatment strategies.

Employment

North Carolina State University 2016 - present
Assistant Professor in Statistics

Harvard University 2014 - 2016
Postdoctoral Fellow in Biostatistics
Research: "Develop causal inference methods with application to initiating ART in HIV-positive patients"
Adviser: Judith J. Lok, jlok@hsph.harvard.edu

Education

Iowa State University, USA 2009–2014
Ph.D. Co-major in Statistics and Applied Mathematics. GPA: 4.0/4.0
Thesis: "Fractional imputation methods in missing data analysis and spatial statistics"
Advisers: Jae Kwang Kim, jkim@iastate.edu and Zhengyuan Zhu, zhuz@iastate.edu

Beijing Normal University, P.R. China 2005–2009
B.Sc. in Mathematics and Applied Mathematics

Awards & Honors

1. **Ralph E. Powe Junior Faculty Enhancement Award**, 2018, Oak Ridge Associated Universities (ORAU).
2. **Research and Innovation Seed Funding**, 2018, North Carolina State University, to assist in developing innovative interdisciplinary programs.
3. **Young Investigator Scholarship**, 2016, Conference on Retroviruses and Opportunistic Infections.
4. **Harvard Postdoctoral Association Travel Award**, 2015, Harvard T. H. Chan School of Public Health, conference travel award for postdoctoral fellows.

5. **American Statistical Association (ASA) Edward C. Bryant Scholarship Award**, 2014, Westat, award for an outstanding graduate student in survey statistics.
6. **Student Paper Competition Award**, 2014, Joint Statistical Meeting (JSM), sponsored by the Social Statistics/Government/Survey Research Methods sections of the ASA.
7. **Research Excellence Award**, 2013, Iowa State University, award for outstanding research by graduate students.
8. **Bancroft Award in Statistics**, 2012, Iowa State University, award to recognize the top student in the doctoral co-major.

Research Grant

1. NSF (National Science Foundation) grant DMS 1811245, 2018–2021, \$120,000, role: PI. Theory and Methods for Causal Inference in Chronic Diseases.
2. NCSU Research and Innovation Seed Funding, 2018–2019, \$31,500, role: PI. Statistical Methods for Oral Anticoagulation Therapy in Patients with Atrial Fibrillation.
3. ORAU Ralph E. Powe Junior Faculty Enhancement Award, 2018–2019, \$10,000, role: PI. Statistical Methods for Comparative Effectiveness Research in HIV infection.
4. NCI (National Cancer Institute) grant P01 CA142538, role: co-Investigator. Statistical Methods for Cancer Clinical Trials.

Publications

1. **S. Yang** and P. Ding (2019). Combining multiple observational data sources to estimate causal effects, *Journal of American Statistical Association*, Doi:10.1080/01621459.2019.1609973. [[arxiv](#)]
2. **S. Yang**, L. Wang, and P. Ding (2019). Nonparametric identification of causal effects with confounders subject to instrumental missingness, *Biometrika*, accepted. [[arxiv](#)]
3. **S. Yang**, K. Pieper, and F. Cools (2019). Semiparametric estimation of structural failure time model in continuous-time processes, *Biometrika*, accepted. [[arxiv](#)]
4. **S. Yang** and D. Zeng (2018). Discussion on penalized spline of propensity methods for treatment comparison by Zhou, Elliott and Little, *Journal of American Statistical Association*, **114**, 30–32.
5. **S. Yang** and J. J. Lok (2018). Sensitivity analysis for unmeasured confounding in coarse structural nested mean models, *Statistica Sinica*, **28**, 1703–1723. [[Link](#)]
6. **S. Yang** (2018). Propensity score weighting for causal inference with clustered data, *Journal of Causal Inference*, doi.org/10.1515/jci-2017-0027. [[Link](#)]
7. **S. Yang** and J. K. Kim (2018). Nearest neighbor imputation for general parameter estimation in survey sampling, *Advances in Econometrics*, **39**, 211–236. [[arxiv](#)]
8. **S. Yang** and P. Ding (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores, *Biometrika*, **105**, 487–493. [[Link](#)]
9. Z. Wang, J. K. Kim, and **S. Yang** (2018). An approximate Bayesian inference under informative sampling, *Biometrika*, **105**, 91–102. [[Link](#)]
10. J. J. Lok, **S. Yang**, B. Sharkey, Hughes, M (2018). Estimation of the cumulative incidence function under multiple dependent and independent censoring mechanisms, *Lifetime Data Analysis*, **24**, 201–223. [[Link](#)]
11. **S. Yang**, A. A. Tsiatis, and M. Blazing (2018). Modeling survival distribution as a function of time to treatment discontinuation: a dynamic treatment regime approach, *Biometrics*, **74**, 900–909. [[Link](#)]
12. **S. Yang** and J. K. Kim (2017). A semiparametric inference to regression analysis with missing covariates in survey data, *Statistica Sinica*, **27**, 261–285. [[Link](#)]
13. J. K. Kim and **S. Yang** (2017). A note on multiple imputation under complex sampling, *Biometrika*, **104**, 221–228. [[Link](#)]

14. **S. Yang** and J. K. Kim (2017). Discussion: dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial? by X. Xie and X. L. Meng, *Statistica Sinica*, **27**, 1568–1573. [[Link](#)]
15. **S. Yang**, and J. J. Lok (2016). A goodness-of-fit test for structural nested mean models, *Biometrika*, **103**, 734–741. [[Link](#)]
16. **S. Yang**, and J. K. Kim (2016). Fractional imputation in survey sampling: a comparative review, *Statistical Science*, **31**, 415–432. [[Link](#)]
17. **S. Yang**, G. Imbens, Z. Cui, D. Faries and Z. Kadziola (2016), Propensity score matching and stratification in observational studies with multi-level treatments, *Biometrics*, **72**, 1055–1065. [[Link](#)] With R package available "[multilevelMatching](#)".
18. **S. Yang** and J. K. Kim (2016). A note on multiple imputation for method of moments estimation, *Biometrika*, **103**, 244–251. [[Link](#)]
19. **S. Yang** and J. K. Kim (2015). Likelihood-based inference with missing data under missing-at-random, *Scandinavian Journal of Statistics*, **43**, 436–454. [[Link](#)] ** Winner of the 2014 JSM Student Paper Competition Award
20. K. L. Peyer, G. Welk, L. B. Davis, **S. Yang**, and J. K. Kim (2015). Factors associated with parent concern for child weight and parenting behaviors, *Childhood Obesity*, **11**, 269–274. [[Link](#)]
21. **S. Yang** and Z. Zhu (2015). Variance estimation and kriging prediction for a class of non-stationary spatial models, *Statistica Sinica*, **25**, 135–149. [[Link](#)]
22. J. K. Kim and **S. Yang** (2014). Fractional hot deck imputation for robust estimation under item nonresponse in survey sampling, *Survey Methodology*, **40**, 211–230. [[Link](#)]
23. J. K. Kim, Z. Zhu, and **S. Yang** (2013). Improved estimation for June Area Survey incorporating several information, *Proceedings 59th ISI World Statistics Congress, Hong Kong, China*, 199–204. [[Link](#)]
24. **S. Yang**, J. K. Kim and D. W. Shin (2013). Imputation methods for quantile estimation under missing at random, *Statistics and Its Interface*, **6**, 369–377. [[Link](#)]
25. **S. Yang**, J. K. Kim and Z. Zhu (2013). Parametric fractional imputation for mixed models with nonignorable missing data, *Statistics and Its Interface*, **6**, 339–347. [[Link](#)]

Technical Reports

26. S. Chen, **S. Yang**, and J.K. Kim. Nonparametric mass imputation for data integration.
27. X. Mao, Z. Wang, and **S. Yang**. Matrix completion for survey data prediction with multivariate missingness. [[arxiv](#)]
28. L. Dong, **S. Yang**, X. Wang, D. Zeng, J.W. Cai. Integrative analysis of randomized clinical trials with real world evidence studies.
29. W. Li, **S. Yang**, and P. Han. Robust estimation for moment condition models with data missing not at random, *Journal of Statistical Planning and Inference*, revision.
30. A. Larsen, **S. Yang**, A. Rappold, and B. Reich. A spatial causal analysis of wildland fire-contributed PM2.5 using numerical model output. *Journal of American Statistical Association*, revision.
31. **S. Yang**, J. K. Kim, and R. Song. Doubly robust inference when combining probability and non-probability samples with high-dimensional data. [[arxiv](#)]
32. **S. Yang**. Semiparametric efficient estimation of structural nested mean models with irregularly spaced observations. [[arxiv](#)]
33. **S. Yang**, J. K. Kim, and Youngdeok Hwang. Integration of survey and big observational data for finite population inference using mass imputation. *International Statistical Review*, revision [[arxiv](#)]
34. **S. Yang** and J. K. Kim. Predictive mean matching imputation in survey sampling. [[arxiv](#)]
35. **S. Yang** and Z. Zhu. Semiparametric estimation of spectral density and variogram with irregular observations, *Journal of Statistical Planning and Inference*, revision. [[arxiv](#)]

Thesis

S. Yang (2014). Fractional imputation method of handling missing data and spatial statistics. Iowa State University. [[Link](#)]

Software

1. An R package [multilevelMatching](#) for implementing a novel matching procedure to compare multiple treatments simultaneously from the observational data.
2. An R package [contTimeCausal](#) provides estimation methods for continuous-time structural failure time models (ctSFTM) and continuous-time Cox marginal structural models (ctCoxMSM).
3. An R package [IntegrativeCI](#) implements integrative analyses for the average treatment effect combining big main data and smaller validation data.

Presentations

Atlantic Causal Inference Conference (ACIC); Eastern North American Region of International Biometric Society Spring Meeting (ENAR); International Chinese Statistical Association (ICSA); Joint Statistical Meeting (JSM)

1. Semiparametric Estimation of Continuous-Time Structural Failure Time Model. *JSM, Denver, Colorado, USA. (Invited) July 2019*
2. JASA, Applications and Case Studies, discussant, *JSM, Denver, Colorado, USA. (Invited) July 2019*
3. Integrative analysis of randomized clinical trial with real world evidence studies. *ICSA, Tianjin, Hebei, China. (Invited) July 2019*
4. Integrative analysis of randomized clinical trial with real world evidence studies. *Real-world analytics team, Eli Lilly and Company, Skype Meeting, USA. (Invited) June 2019*
5. Causal inference with confounders missing not at random. *ICSA, Raleigh, NC, USA. (Invited) June 2019*
6. Causal inference with confounders missing not at random. *Departmental seminar speaker, Raleigh, NC, USA. April 2019*
7. Causal inference with confounders missing not at random. *ENAR, Philadelphia, PA, USA. (Invited) March 2019*
8. Causal inference with confounders missing not at random. *Colloquium Speaker, Baylor University, Waco TX, USA. (Invited) February 2019*
9. Propensity score matching and subclassification in observational studies with multi-level treatments. *The International Biometrics Society Journal Club. Webinar (Invited) December 2018*
10. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity score. *ICSA, New Brunswick, NJ, USA. (Invited) June 2018*
11. Combining multiple observational data sources to estimate causal effects. *ACIC, Pittsburgh, PA, USA. (Invited) May 2018 [[Abstract](#)]*
12. Dynamic regime marginal structural models to survival distribution as a function of time to treatment discontinuation. *ENAR, Atlanta, GA, USA. (Invited) March 2018 [[Abstract](#)]*
13. Dynamic regime marginal structural models to survival distribution as a function of time to treatment discontinuation. *Colloquium Seminar Speaker, Kansas State University, Manhattan, KS, USA. March 2018 [[Abstract](#)]*
14. Modeling survival distribution as a function of time to treatment discontinuation. *Departmental Seminar Speaker, North Carolina State University, Raleigh, NC, USA. September 2017 [[Abstract](#)]*

15. Modeling survival distribution as a function of time to treatment discontinuation. *Colloquium Speaker, Purdue University, West Lafayette, IN, USA. (Invited) September 2017* [[Abstract](#)]
16. Estimation of the cumulative incidence function under multiple dependent and independent censoring mechanisms. *JSM, Baltimore, MD, USA. (Invited) August 2017* [[Abstract](#)]
17. Nonparametric identification of causal effects with confounders subject to instrumental missingness. *ACIC, Chapel Hill, NC, USA. (Poster) May 2017* [[Poster](#)]
18. Propensity score weighting for causal inference with multi-stage data. *ENAR, Washington, DC, USA. (Invited Poster) March 2017* [[Abstract](#)] [[Poster](#)]
19. A note on multiple imputation of handling missing data under complex sampling. *JSM, Chicago, IL, USA. (Invited) August 2016* [[Abstract](#)]
20. Estimation and goodness-of-fit test of structural nested mean models. *ENAR, Austin, TX, USA. March 2016* [[Abstract](#)]
21. Optimal estimation of coarse structural nested mean models with application to initiating HAART in HIV-positive patients. *National Institute of Health (NIH) Infectious Disease Research: Quantitative Methods and Models in the era of Big Data Statistical Workshop, Bethesda, MD, USA. (poster) Nov 2015*
22. Double robust goodness-of-fit test of coarse structural nested mean models with application to initiating HAART in HIV-positive patients. *ACIC, Philadelphia, PA, USA. May 2015*
23. Optimal estimation of coarse structural nested mean models. *ENAR, Miami, FL, USA. March 2015* [[Abstract](#)]
24. Fractional imputation method for missing data analysis: a review. *Workshop on Analyzing Complex Survey Data with Missing Item Values, National Institute of Statistical Sciences (NISS), Washington D.C., USA. (Invited) October 2014* [[Abstract](#)] [[Slides](#)]
25. Likelihood-based inference with missing data under missing-at-random. *JSM, Boston, MA, USA. August 2014* [[Abstract](#)]
26. Likelihood-based inference with missing data under missing-at-random. *ICSA-KISS symposium, Portland, OR, USA. June 2014* [[Abstract](#)]
27. Propensity score matching and subclassification with multivalued treatments. *ACIC, Providence, RI, USA. May 2014* [[Abstract](#)]
28. Fractional hot deck imputation for robust estimation under item nonresponse in survey sampling. *JSM, Montreal, Canada. August 2013* [[Abstract](#)]
29. Parametric fractional imputation for longitudinal data with non-ignorable missing data. *JSM, San Diego, CA, USA. August 2012* [[Abstract](#)]
30. Quantify uncertainty in image classification. *National Agricultural Statistics Service (NASS), Washington, DC, USA. May 2011*

Professional Experience

1. Organize an invited session on “Recent Advance of Causal Inference in Failure Time Settings” at *JSM, Colorado, Denver, USA. August 2019*
2. Organize an invited session on “Causal Inference and Missing Data Analysis: Identification and Estimation” at *ICSA, Raleigh, NC, USA. June 2019*
3. Organize an invited session on “Causal Inference with Non-ignorable Missing Data: New Developments in Identification and Estimation” at *ENAR, Philadelphia, PA, USA. March 2019*
4. Organize and chair an invited session on “Statistical Inference in Air Pollution and Health Epidemiology” at *ICSA, New Brunswick, NJ, USA. June 2018* [[Abstract](#)]
5. Organize and chair an invited session on “Causal Inference and Data Fusion” at *ACIC, Pittsburgh, PA, USA. May 2018* [[Abstract](#)]
6. Organize and chair a session on “Causal Inference for Continuous-time Processes: New Developments” at *ENAR, Washington, DC, USA. March 2017* [[Abstract](#)]

7. Organize an invited session on “New Developments in Structural Nested Models with Medical Applications” at *ACIC, Philadelphia, PA, USA. May 2015* [[Abstract](#)]
8. Organize an invited session on “Big Data Techniques for Survey Data Integration” at *JSM, Seattle, WA, USA. August 2015* [[Abstract](#)]
9. Translate chapters in Book “Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction” into Chinese (2015–2017)

Students Supervision

1. Nathan Corder (PhD Co-Advisor)
2. Lin Dong (PhD Co-Advisor)
3. Alexandra Larsen (PhD Committee Member)
4. Qian Guan (PhD Committee Member)
5. Breanna Swan (PhD Committee Member)
6. Eric Rose (PhD Committee Member)

Editorial Position

1. Associate Editor, *Statistica Sinica* 2016 – present
2. Associate Editor, *Biometrics* 2018 – 2020

Journals Reviewed

1. NSF DMS panel review
2. Annals of Applied Statistics
3. Biometrika
4. Biometrical Journal
5. Canadian Journal of Statistics
6. Communications in Statistics
7. Epidemiologic Methods
8. Electronic Journal of Statistics
9. International Journal of Biostatistics
10. Journal of Business & Economic Statistics
11. Journal of Multivariate Analysis
12. Journal of the American Statistical Association
13. Journal of the Korean Statistical Society
14. Journal of the Royal Statistical Society: Series B
15. Statistica Sinica
16. Stat
17. Statistics and Its Interface
18. Survey Methodology